Considerations in Selecting and Using Measures for RCT’s

William F. Chaplin, Ph. D.
Department of Psychology
St. John’s University
New York

“All models are wrong. Some are useful”

------------attributed to George Box

“All measures are wrong. Some are useful”

---------An anonymous psychometrician

Measure = Model

''Measurement is the Achilles' heel of socio-behavioral research. Although most programs in socio-behavioral sciences ... require a medium of exposure to statistics and research design, few seem to require the same where measurement is concerned ...

It is, therefore, not surprising that little or no attention is given to the properties of the measures used in many research studies.''

-------Pedhazur, E.J., & Schmelkin, L.P. (1991) Measurement, Design and Analysis: An Integrated Approach. Hillsdale NJ: Erlbaum. (p. 2-3).

“Yet, when we teach courses on measurement and test construction, we seldom encounter much enthusiasm. In fact, most students think that measurement is outright boring.”

------John & Benet-Martinez (2000)

Measurement and RCTs

RCT’s are NOT about measurement. They are about:

1) For what (if anything) does the treatment work? (Dependent variable)
2) Why does it work? (Mediating variable)
3) For whom does it work? (Moderating variable)

Obviously we need MEASURES of all these variables, but the measures are a means to an end.

As the lectures have already made clear, we have a great deal to worry about in designing, conducting, and interpreting a RCT. Adding measurement selection to the list is a distraction we would like to avoid.

Nonetheless, as much as we would like to just pull a measure off the shelf and get on with recruitment, in my two lectures I intend to increase your awareness of measurement issues and enhance your appreciation of why you should care at least a little bit about these issues.

Measures should be reliable and valid.

“*Reliability coefficients for X are reported to range from .70 to .85 (ref). According to the manual (ref) X has been found to be a valid measure of A.*” --- Anon

Issue 1

Reliability and Validity are not Immutable Properties of Measures, they will vary as a function of sample, situation, and the purpose of measurement.

A measure of self-esteem that is valid for college students may not be as valid in a sample of 70 year old cardiac rehabilitation patients.

A measure of whole blood serotonin that is reliable (stable) in adults may not be stable in young children.

There is no such thing as “a reliable measure” or “a valid measure.”

Formally we study how measures behave differently in different groups or situations by assessing DIF “Differential Item Functioning” using IRT (Item Response Theory) models.

A personally distressing example

Juror Bias Scale

Kassin, S. M., & Wrightsman, L. S. (1983). The construction and validation of a juror bias scale. *Journal of Research in Personality, 17,*423-442.

17 items; split half reliability = .81 in a sample of 221 college students in a mock jury setting

Clark and Chaplin

Administered the juror bias scale to 388 individuals who were serving in a jury pool in Alabama to see, among other things, if bias was related to selection.

But our estimate of split-half reliability for this scale in this sample is .35. (!)

We are now writing a far different, and less exciting manuscript.

Recommendations:

> Select measures that have been evaluated on samples and in situations similar to your RCT

> Evaluate the measure on your data (most easily done with coefficient alpha, but alpha is not always appropriate see below)

>This means you must enter the items, not just the total score in your data set.

>Do not assume that a measure assesses the same construct in the same way in different groups.

Especially important in subgroup analysis !

Issue 2

Reliability is an internal property of a measure. It concerns how much of the variability in a measure (across people) is systematic (replicable) as opposed to random.

$\sigma_X^2 = \sigma_T^2 + \sigma_E^2$

reliability = σ_t^2 / σ_X^2

Reliability is usually estimated in two ways.

stability (correlations between same measure at two times)

internal consistency (coefficient alpha)

Neither of these necessarily estimates reliability

Stability (“test-retest reliability”) only assesses reliability if a measure is expected to be stable (except for random error)

We do not expect stability on “state” measures such as the state anxiety scale

Internal consistency only assesses reliability if responses to items are expected to be consistent (except for random error).

We do not expect consistency in mutually exclusive behavioral coding systems where doing one behavior means not doing another.

What you may not know about coefficient alpha.

1. It is NOT a reliability coefficient in the classic sense. It is a generalizability coefficient. (Influenced by both measurement error AND item content.)

2. Alpha is NOT a measure of a scales unidimensionality, rather the legitimacy of alpha as a psychometric measure assumes unidimensionality

3. Alpha depends upon the number of items a scale contains as well as their consistency.

4. A high alpha can “paradoxically” reduce a scale’s validity. Called the “attenuation paradox.”

(See assigned Chapter by John and Benet-Martinez)

A formula used to calculate alpha is:

$$\frac{k * \mathrm{avg}(r_{ij})}{[1+ \quad (k-1)*\mathrm{avg}(r_{ij}]}$$

where ***k*** is the number of items and ***avg(r_{ij})*** is the mean of the inter-item correlations.

Note that alpha depends on both the number of items and the correlations among them. Even when the average correlation is small, the reliability coefficient can be large if the number of items is large.

Example: Attributional Style Questionnaire

The average inter-item correlation for the internal subscale scale was .10 and the original number of items was 10. What was the internal consistency.

$10(.10) / [1 + 9(.10)] = .53$ not good.

Increase the number of items to 15

$15(.10)/[1 + 14(.10)] = .63$ better

Increase the number of items to 25

$25 (.10) / [1 + 24(.10)] = .74$ “acceptable”

Increase the number of items to 40

$40(.10) / [1 + 39(.10)] = .82$ “good”

NOTE: The additional items must come from the same domain “universe” as the original items.

Individual Items are mostly “noise” or error. We increase the “signal” of the underlying construct (tendency to make internal or external attributions) to the random noise (item idiosyncrasy) of error through the POWER of AGGREGATION

Cook Medley Hostility Scale Example

Data from subset (N = 600) of HOT Study Participants

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.884	.885	50

k

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Inter-Item Correlations	.134	-.127	.583	.709	-4.605	.009	50

The covariance matrix is calculated and used in the analysis.

Average r_{ii}

Item Evaluation

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation		Cronbach's Alpha if Item Deleted
host1	17.72	72.982	.218	.	.883
host2	17.69	72.503	.271	.	.883
host3	17.56	71.364	.366	.	.881
host4	17.29	73.235	**.115**	.	.885
host5	17.68	72.486	.269	.	.883
host6	17.36	73.239	**.113**	.	.885
host7	17.57	71.193	.392	.	.881
host8	17.33	73.635	**.067**	.	.886
host9	17.75	72.879	.260	.	.883
host10	17.57	71.702	.328	.	.882
host11	17.70	73.195	.172	.	.884
host12	17.80	73.171	.273	.	.883
host13	17.66	70.997	.479	.	.880
host14	17.71	71.590	.436	.	.881
host15	17.35	74.230	**-.003**	.	.887
host16	17.50	70.007	.519	.	.879
host17	17.74	71.836	.437	.	.881
host18	17.66	70.751	.515	.	.879
host19	17.52	70.760	.431	.	.880
host20	17.47	72.366	.222	.	.883

My ways of doing things are apt to be misunderstood by others

I am quite often not in on the gossip or talk of the group I belong to.

I am against giving money to beggars

I frequently ask people for advice

My relatives are nearly all in sympathy with me

Remember the attenuation paradox !

Issue 3

Establishing a measures validity is a complex and on-going process. It is equivalent to testing a theory (in this case a measurement theory) so it is never completed.

As our understanding (theories) of constructs (e.g. blood pressure, anxiety, depression) evolve, our measures (representations of those constructs) *should* evolve.

The measure of the thing is not the thing

(except maybe length)

This limitation of measurement results in

- “Operational definitions” or “gold standard” (a royal cop out)

“Intelligence is whatever intelligence tests measure”

-------Boring, E.G. (1923) Intelligence as the tests test it. *The New Republic*, June, 35-189.

- Historical precedent as the basis for measurement (not all bad)

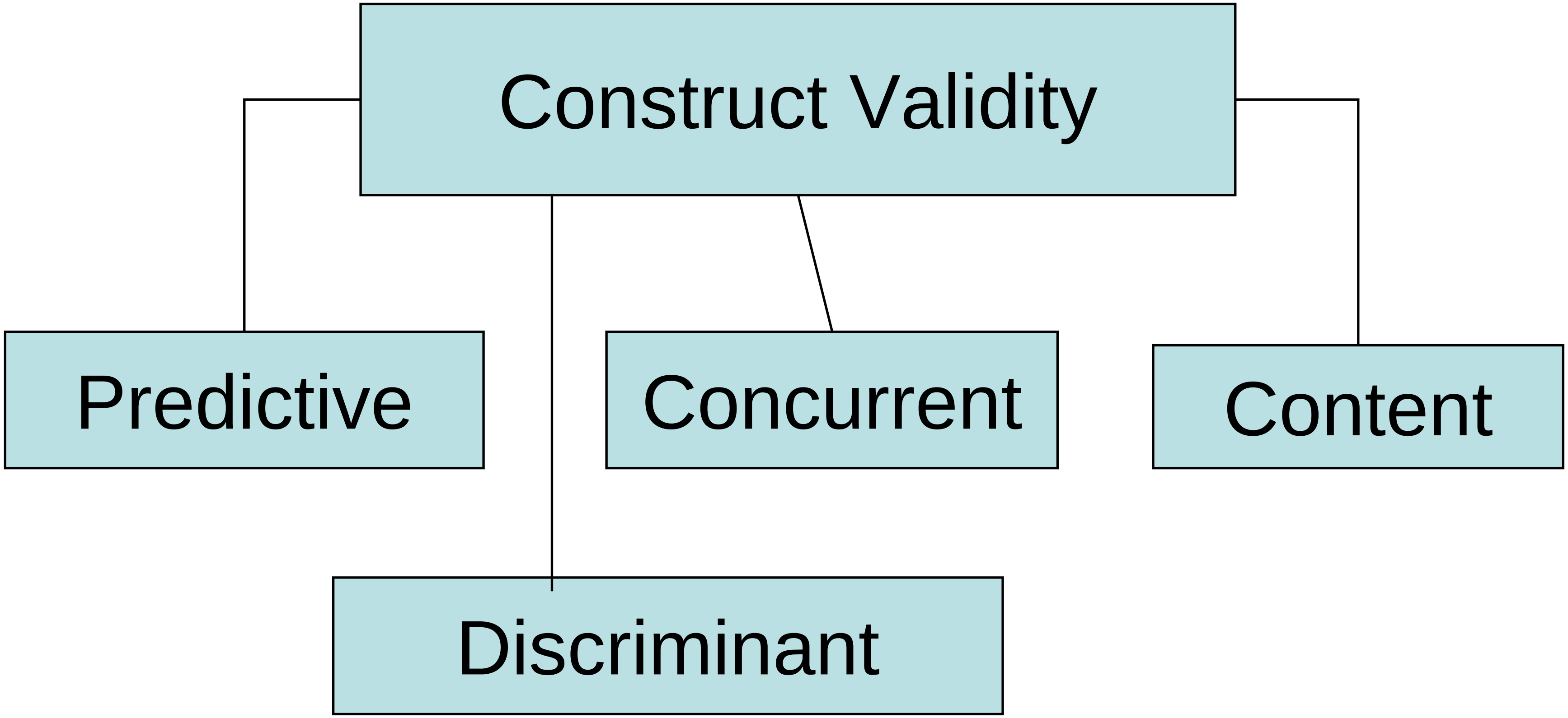
- Fear of measurement

Validity and Type III error

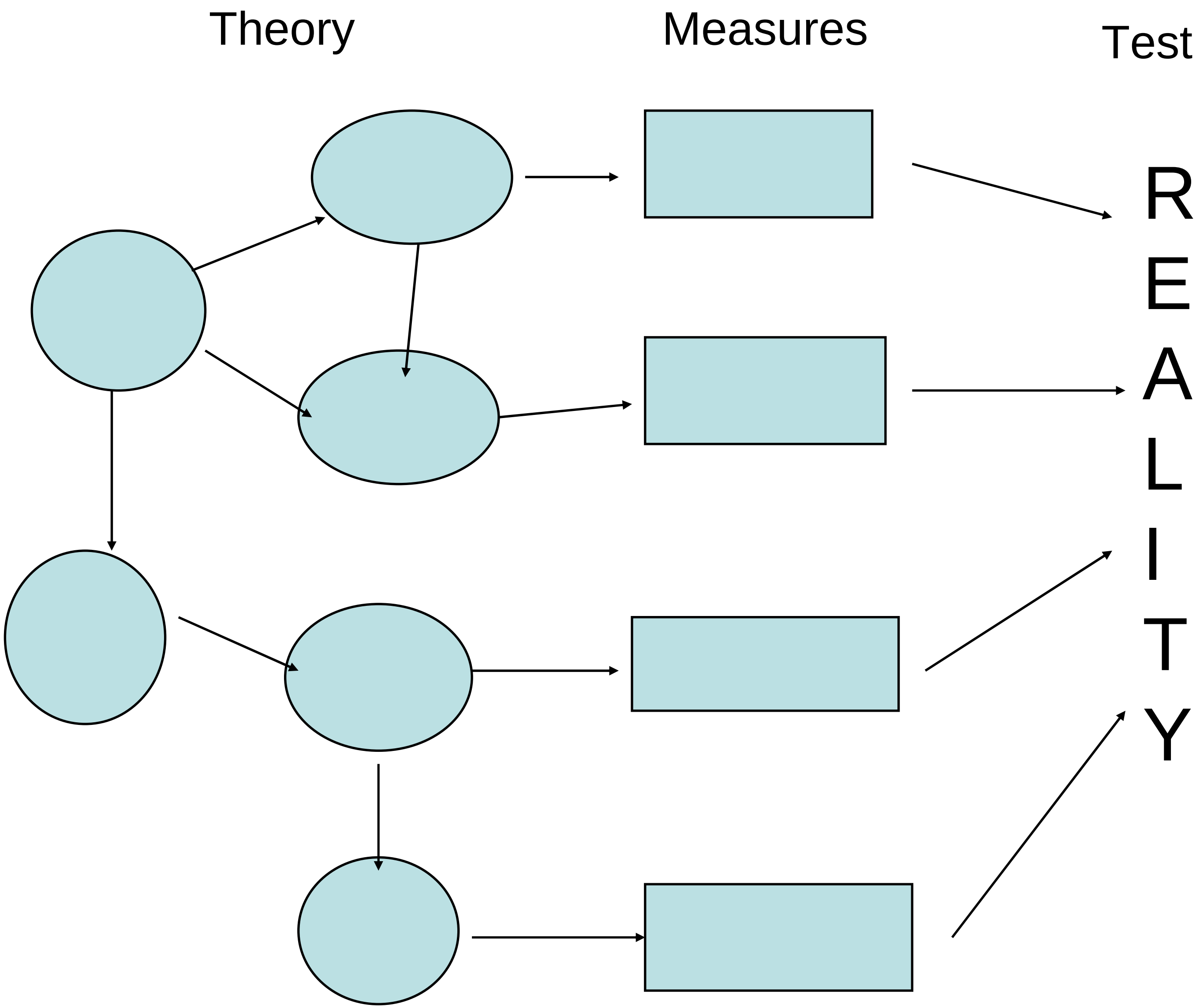
Type III error has been used to refer to a variety of mistakes, but I am using it here to refer to “solving the wrong problem correctly” or “correctly answering the wrong question.”

In this case the wrong question is caused by using an invalid measure: A measure that represents something other than what we think it does.

Validity


Construct Validity
Predictive
Concurrent
Content
Discriminant

Nomological Network Cronbach and Meehl (1955)



Circularity of Substantive and Measurement Theories

1) X is a measure of A
2) Y is a measure of B
3) A and B are negatively correlated
4) Are X & Y negatively correlated?

Hopefully the circularity is an upward spiral!

Content Validity

Item Content Provides an Ostensive Definition of a Construct

There is no good quantitative index of content validity

But content can provide clues to interpreting study results. Did we measure what we thought we did?

Loneliness as an Example

“Because most of the self report measures [of loneliness] for children contain diverse item content that goes beyond loneliness per se (as does the widely used UCLA Loneliness Scale for adults), caution must be used when interpreting the results. Some investigators have therefore calculated ‘pure loneliness’ scores by using only items that directly assess feelings of loneliness.”

Asher, S. R. & Paquette, J. A, (2003) Loneliness and peer relations in children. *Current Directions in Psychological Science, 12,* 75-78.

UCLA Loneliness Scale (Version 3)

Russell, D. W. (1996). Loneliness scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment, 66,* 20-40.

Response is on a 4 point scale 1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Always.

How often do you feel…..

1) that you are in tune with the people around you?

2) that you lack companionship?

3) that there is no one you can turn to?

4) alone?

5. part of a group of friends?
6. that you have a lot in common with the people around you?
7. that you are no longer close to anyone?
8. that your interests and ideas are not shared by those around you?
9. outgoing and friendly?
10. close to people?
11. left out?
12. that your relationships with others are not meaningful?
13. that no one really knows you well?
14. isolated from others?

15. that you can find companionship when you want it?
16. that there are people who really understand you?
17. shy?
18. that there are people around you but not with you?
19. that there are people you can talk to?
20. that there are people you can turn to?

NYU Loneliness Scale

Rubenstein and Shaver (1982)

I am a lonely person.

How often do you feel lonely

I will always be a lonely person

Other people think of me as lonely

I always was a lonely person

Compared to other people how lonely do you think you are?

When I am completely alone, I feel lonely?

When you feel lonely how lonely do you feel?

Loneliness Scale

de jong, G. J., & Kamphuis, F. H. (1985). The development of a Rasch-type loneliness scale. *Applied Psychological Measurement, 9,* 289-299.

1. There is always someone I can talk to about my day to day problems.
2. I miss having a really close friend.
3. I experience a general sense of emptiness.
4. There are plenty of people I can lean on when I have problems.
5. I miss the pleasure of the company of others

6. I find my circle of friends too limited.

7. There are many people I can trust completely.

8. There are enough people I feel close to.

9. I miss having people around.

10. I often feel rejected.

11. I can call on my friends whenever I need them.

Differential Loneliness Scale

Schmidt, N., & Sermat, V. (1983). Measuring loneliness in different relationships. *Journal of Personality and Social Psychology, 44,* 1038-1047.

Family

I feel close to members of my family.

I have little contact with members of my family

I do not get along well with members of my family

I have a good relationship with most members of my immediate family

My family seldom really listens to what I say.

Romantic/Sexual

I have a lover or spouse with whom I can discuss my important problems and worries

I am now involved in a romantic or marital relationship where both of us are making a genuine effort at cooperation

My lover or spouse sense when I am troubled or encourages me

I feel valued and respected in my current romantic or marital relationship.

I seldom get the emotional security I need from a good romantic or sexual relationship

Friends

I do not feel that I can turn to my friends living around me for help when I need it.

I allow myself to become close to my friends

I do not have many friends in the city where I live

I get plenty of help and support from my friends

Few of my friends understand me the way I want to be understood.

Groups/Community

I feel I really do not have much in common with the larger community in which I live.

No one in the community where I live seems to care much about me

I feel that I have “roots” (sense of belonging) in the larger community or neighborhood I live in.

I do not have any neighbors who would help me out in a time of need

I know people in my community who understand and share my views and beliefs

Summary

I am not convinced that any of these measures provide a pure index of loneliness.

Social support

Shyness

Introversion

Neuroticism

Predictive Validity: A perspective.

The most common way we evaluate a measures validity is to determine that the measure predicts other measures it should (*according to our theories of the constructs!!!*)

1. "Good" predictive validity does not PROVE that the measure is valid. Poor predictive validity provides stronger evidence against the measure. As always a disconfirming result is stronger than a confirming one.

2. What do we mean by "good?"

Parlor Games

What is the relation between…...

1. gender and risk-taking behavior? (Males are higher)

2. Exposure to media violence and interpersonal aggression?

3. Prominent movie critics ratings and box office success?

4. Gender and weight?

5. Anti-hypertension medication and stroke?

6. Nearness to equator and daily temperature?

What is the relation between…...

1. gender and risk-taking behavior? (Males are higher)

r = .09

2. Media violence and interpersonal aggression?

r = .13

3. Prominent movie critics ratings and box office success?

r = .17

4. Gender and weight?

r = .26

5. Taking antihypertension medication and stroke?

r =-.03 (N = 59,086)

6. Nearness to equator and daily temperature?   r = .60

Predictive validity of psychological and medical measures and consequential outcomes

1. Fecal blood test screening and reduced death from colorectal cancer r = .01 (N = 329,624)

2. Beck hopelessness scale scores and subsequent suicide

r = .08

3. Single serum progesterone testing and diagnosis of ectopic pregnancy

r = .23

4. Jenkins Activity Survey scores and heart rate and blood pressure activity.

r = .26

Recommendations for Measurement in RCT

1. Spend some time reviewing evidence for the measures you plan to use.

In this lecture I have tried to broaden your perspectives on measurement so that you can approach this task thoughtfully

2. If there are no acceptable measures available you may want to re-think your study. I do not recommend either:

a) Using a poor quality measure, or

b) Developing your own measure (RCT's should not be measurement studies!!)

Point for Discussion:

There is sometimes a strong tension between using a good measure (or developing one) and historical precedent based on what may be poor measures (e.g. MD assessed “office blood pressures” as the “gold standard” in hypertension research; the Cook-Medley Hostility Scale).

The issue is the desire to relate our research to previous research for the sake of the continuity of science (if not for something so crass as being able to publish one’s findings!). But continuity of misleading results based on wretched measures may not be a service to those whose lives and health we are trying to improve.